Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms

Erwin Lanceta Cahapin[1], Beverly Ambagan Malabag[1], Cereneo Sailog Santiago Jr.[1], Jocelyn L. Reyes[2], Gemma S. Legaspi[3], Karl Louise Adrales[4]
[1]Department of Information Technology, Cavite State University–Silang Campus, Cavite, Philippines
[2]Department of Teacher Education, Cavite State University–Silang Campus, Cavite, Philippines
[3]College of Arts and Sciences, Cavite State University–Indang Campus, Cavite, Philippines
[4]Department of Arts and Sciences, Cavite State University–Silang Campus, Cavite, Philippines

Article Info

Article history:

Received Sep 23, 2022
Revised Apr 17, 2023
Accepted May 3, 2023

Keywords:

Clustering
Density-based spatial clustering of applications with noise algorithm
Hierarchichal
K-means
University admission

ABSTRACT

Admissions in the university undergo procedures and requirements before a student can be officially enrolled. The senior high school grades remain the most significant in college admission decisions. This paper presents the use of data mining to cluster students based on admission datasets. The admission dataset for 2019-2020 was obtained from the office of student affairs and services. This dataset contains 2,114 observations with 11 attributes. Data preparation and data standardization were performed to ensure that the dataset is ready for processing and implemented in R programming language. The optimal number of clusters (k) was identified using the silhouette method. This method gave an optimal number of k=2 which was used in the actual clustering using the k-means and hierarchical clustering algorithms. Both algorithms were able to cluster students into two: cluster 1-social sciences or board courses and cluster 2-management or non-board courses. Further, density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm was also used on the same dataset and it yielded a single cluster. This study can be replicated by using at least a 5-year dataset of students' admission data employing other algorithms that would suggest students' retention and turn over to board examinations.

Corresponding Author:

Cereneo Sailog Santiago Jr.
Department of Information Technology, Cavite State University–Silang Campus
Silang, Cavite, Philippines
Email: cssantiago@cvsu.edu.ph

1. INTRODUCTION

In developing countries, improving the education sector is the principal interest of every government institution. The main role of higher education institutions (HEIs) is to attain global competitiveness and produce skilled human capital resources. Keeping standards in check is strenuous for any institution in line with the latest labor market demands [1]. The implementation of the k-12 program has drastically changed Philippine education perspectives and HEIs have been facing difficulties admitting freshmen applicants and incoming enrollees [2].

First-year college students in the Philippines are placed in their courses based on their senior high school strand, grade point average (GPA), and entrance examination results. Meeting these criteria will ideally lead students to the course that best suits them. However, failure to meet any of these criteria will cause a course mismatch, such as a low GPA will deny entry to board courses, low exam results will deny entry to board courses, and strand mismatch is not permitted. In these circumstances, clustering and cluster analysis may be able to provide insights to admission officers as to where students will be placed. The use of data mining techniques can help overcome the difficulties of forecasting students' enrollment preferences.

HEIs monitor which students need to qualify for particular programs and which students need further assistance to be eligible to graduate. However, this way is mostly difficult for HEIs to track the path of students. A viable way to counter these difficulties is through data analysis and data mining [3]. Data mining is a crucial method of distinguishing and deciphering information from a large quantity of data. This innovation is essential for academic institutions to work with greater efficiency. Using data mining techniques, student results can be foreseen. The substantive feature of data mining is to analyze a vast quantity of data intending to separate obscure patterns, such as cluster analysis, anomaly detection, and association rule mining [4].

Students desire to earn a degree from a reputable university and believe that finding the right advice on which programs to enroll will not be an issue [5]. During college registrations, students undergo an entrance examination to qualify for their desired program. However, college programs often lack appropriate information, causing students to be less concentrated in college. Students browse for information on their chosen program through college websites because there is no system for students to determine their program. Although, there is a decision support system provided to help students in their decision. This decision support system is supported by density-based spatial clustering of applications with noise (DBSCAN). This method is used for clustering and grouping of available data [6].

This study performed clustering of first-year college students using admission dataset as implemented in R and compared the results of the different clustering algorithms: k-means, hierarchical, and DBSCAN. When seeking more effective technology to better manage and support decision-making processes or helping academic institutions to develop new policies and plan for better management of the current processes [7], clustering admission data will increase the quality of their administrative decisions as they establish effective admissions standards [8]. Moreover, creating new techniques for information discovery from databases used in education systems may be applied to better decision-making [5], [9].

2. LITERATURE REVIEW

Clustering is an unsupervised statistical data analysis technique where data is divided into subsets called clusters to find useful and hidden patterns [10], [11]. Additionally, cluster analysis allows a data scientist to look at the data from a different perspective without preconceived profiles [12]. In this paper, the terms clustering and cluster analysis are used interchangeably. Several techniques, including k-means, hierarchical, and DBSCAN, can be used to do cluster analysis [10]. One of the most used algorithms is k-means, which employs Euclidean distance as dissimilarity measure, tries to minimize within cluster distance, and maximize between-cluster distance [10], [13]. The k-means clustering separates a dataset D of n items into k groups after receiving the input of dataset D and the parameter k. The resultant intra cluster similarity is strong, while the inter cluster similarity is low because this split depends on the similarity measure. Cluster similarity is calculated using the mean value of the components in a cluster, which can be shown as the cluster's mean [14]. On the other hand, hierarchical clustering joins items to create clusters based on the presence of similar qualities. Hierarchical clustering refers to a process in which clusters in a hierarchy merge with one another at specific distances [15]. It creates a hierarchy to decide cluster allocations. This is accomplished using either a top-down or bottom-up methodology. A dendrogram, which is a point hierarchy based on a tree, is the end product of these procedures [14]. Lastly, DBSCAN is a clustering algorithm that can be used with datasets containing noise points. It can pinpoint noise points and exclude them from the results [16]. By examining the local density of corresponding items, DBSCAN is able to locate clusters in a huge spatial dataset. The DBSCAN algorithm has an edge over the k-means technique in that it can identify data points that are noise or outliers. DBSCAN can locate locations that do not belong to any clusters. However, it still scales to quite big datasets while being slower than agglomerative clustering and k-means [14].

K-means clustering was employed by [17] in one of their studies to look at school entry data. It was demonstrated to be efficient in identifying patterns and trends in the data and may be helpful in guiding admissions decision-making. The authors assert that educational institutions may modify their admission requirements to consider the traits of successful applicants. To forecast students' admission status to higher education based on their academic performance and other criteria, Santosa *et al.* [18] found that k-means clustering is effective in predicting admission status and can be used by schools to identify the students who are most likely to succeed in their programs. The admission process may be improved by classifying applicants into groups based on their traits and suitability [19].

Iqbal *et al.* [20] analyzed data on school entry using the hierarchical clustering approach. They used the Ward's linkage method to cluster students based on their academic achievement, the algorithm successfuly identified several groups of students with diverse academic profiles. The algorithm can be used to identify students who need additional support or interventions to succeed academically. Goyal and Vohra [21] used a hierarchical algorithm to analyze admissions data from higher education. They found that by identifying the traits of qualified applicants, hierarchical clustering can enhance the admissions process by successfully classifying applicants into various groups based on their features. Bowers [22] used hierarchical clustering to classify students based on test results. They found that the technique successfully identified diverse student clusters with distinct performance characteristics. The algorithm can be used to identify areas where students may need extra assistance or to adjust admission requirements so that they more accurately identify applicants with the potential to succeed.

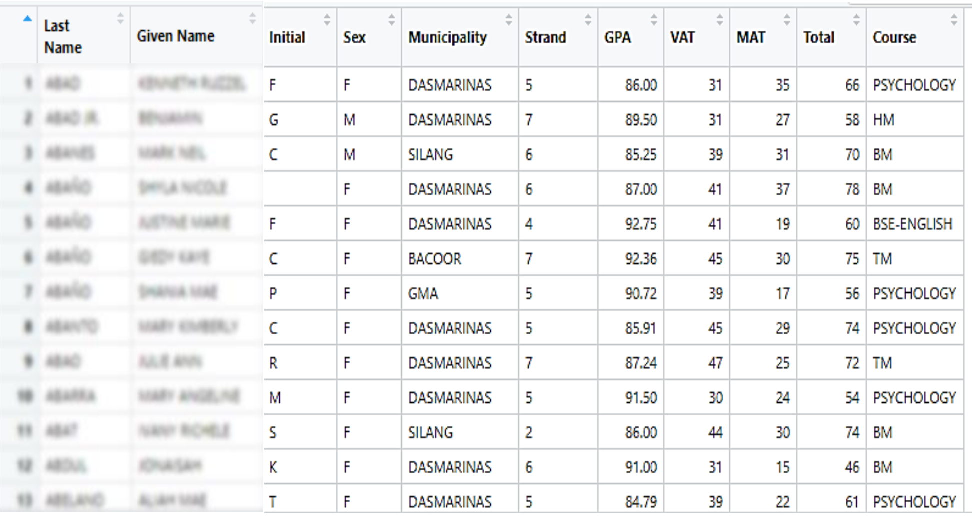
Daniati [23] used the DBSCAN clustering approach to examine school entry data in one study. The study discovered that the algorithm was successful in identifying clusters of applicants with similar features by classifying applicants based on their academic standing and demographic criteria. By identifying the qualities of successful applicants, they suggested applying the algorithm to enhance the admissions process. Nafuri *et al.* [24] examined admission data from higher education using DBSCAN clustering. The algorithm proved successful in separating different groups of candidates based on the criteria when it was used to cluster applicants based on their academic achievement and other characteristics. They believe that by identifying areas where applicants might benefit from additional help, the algorithm may be used to enhance the admissions process. The same method was effectively applied by [25] to identify various student groups with distinctive performance profiles based on the test results of the students. The algorithm might be used to identify areas where students might need extra assistance or to adjust admission requirements, so they better recognize applicants with the potential to succeed.

3. METHOD

Cluster analysis was performed on an admission dataset to identify distinct groups among the students. The following steps were undertaken to perform cluster analysis of the admission dataset:

a. Dataset description

The admission dataset for academic year 2019-2020 was obtained from the Office of Student Affairs and Services of Cavite State University-Silang Campus. The dataset comprises 2,114 observations with 11 attributes. Seven of these attributes are categorical (last name, given name, MI, sex, strand, municipality, and course) while 4 attributes are numerical (GPA, VAT, MAT, and total). Figure 1 shows a portion of the dataset, though some information has been blurred for data privacy.



	Last Name	Given Name	Initial	Sex	Municipality	Strand	GPA	VAT	MAT	Total	Course
1	[redacted]	[redacted]	F	F	DASMARINAS	5	86.00	31	35	66	PSYCHOLOGY
2	[redacted]	[redacted]	G	M	DASMARINAS	7	89.50	31	27	58	HM
3	[redacted]	[redacted]	C	M	SILANG	6	85.25	39	31	70	BM
4	[redacted]	[redacted]		F	DASMARINAS	6	87.00	41	37	78	BM
5	[redacted]	[redacted]	F	F	DASMARINAS	4	92.75	41	19	60	BSE-ENGLISH
6	[redacted]	[redacted]	C	F	BACOOR	7	92.36	45	30	75	TM
7	[redacted]	[redacted]	P	F	GMA	5	90.72	39	17	56	PSYCHOLOGY
8	[redacted]	[redacted]	C	F	DASMARINAS	5	85.91	45	29	74	PSYCHOLOGY
9	[redacted]	[redacted]	R	F	DASMARINAS	7	87.24	47	25	72	TM
10	[redacted]	[redacted]	M	F	DASMARINAS	5	91.50	30	24	54	PSYCHOLOGY
11	[redacted]	[redacted]	S	F	SILANG	2	86.00	44	30	74	BM
12	[redacted]	[redacted]	K	F	DASMARINAS	6	91.00	31	15	46	BM
13	[redacted]	[redacted]	T	F	DASMARINAS	5	84.79	39	22	61	PSYCHOLOGY

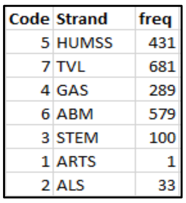
Figure 1. Admission dataset

b. Nominal data coding

The strand attribute is represented by a numeric code, which follows the technique presented in [26]. In this technique, elements are assigned cardinal values based on their frequency, with higher frequency elements receiving higher cardinal values. The code is presented in Figure 2.

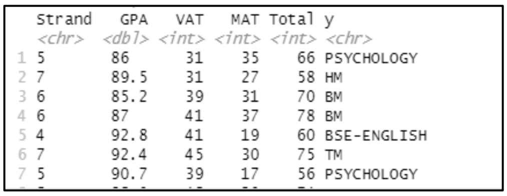
c. Data cleaning

Data cleaning removed the noise data or invalid data. In this step, last name, given name, MI, sex, and municipality attributes were removed. The cleaned dataset is shown in Figure 3.



Code	Strand	freq
5	HUMSS	431
7	TVL	681
4	GAS	289
6	ABM	579
3	STEM	100
1	ARTS	1
2	ALS	33

Figure 2. Codes for the strand attribute

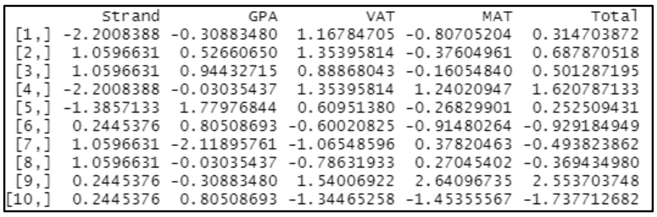


```
  Strand   GPA   VAT   MAT Total y
  <chr>  <dbl> <int> <int> <int> <chr>
1 5       86      31    35    66 PSYCHOLOGY
2 7       89.5    31    27    58 HM
3 6       85.2    39    31    70 BM
4 6       87      41    37    78 BM
5 4       92.8    41    19    60 BSE-ENGLISH
6 7       92.4    45    30    75 TM
7 5       90.7    39    17    56 PSYCHOLOGY
```

Figure 3. Portion of the cleaned data

d. Standardizing the data

The data is transformed into a standardized form. The dataset was standardized to bring the numerical variables to a common scale. Standardization transforms the data by subtracting the mean and dividing by the standard deviation, ensuring that all variables have a similar influence on the clustering algorithm. This step is crucial to avoid any biases introduced by variables with larger scales dominating the clustering results. Figure 4 shows the standardized data.



```
          Strand         GPA         VAT         MAT        Total
 [1,] -2.2008388 -0.30883480  1.16784705 -0.80705204  0.314703872
 [2,]  1.0596631  0.52660650  1.35395814 -0.37604961  0.687870518
 [3,]  1.0596631  0.94432715  0.88868043 -0.16054840  0.501287195
 [4,] -2.2008388 -0.03035437  1.35395814  1.24020947  1.620787133
 [5,] -1.3857133  1.77976844  0.60951380 -0.26829901  0.252509431
 [6,]  0.2445376  0.80508693 -0.60020825 -0.91480264 -0.929184949
 [7,]  1.0596631 -2.11895761 -1.06548596  0.37820463 -0.493823862
 [8,]  1.0596631 -0.03035437 -0.78631933  0.27045402 -0.369434980
 [9,]  0.2445376 -0.30883480  1.54006922  2.64096735  2.553703748
[10,]  0.2445376  0.80508693 -1.34465258 -1.45355567 -1.737712682
```

Figure 4. Portion of the standardized data

e. Identifying the optimal number of clusters for k-means and hierarchical clustering algorithms using silhouette method

The silhouette method was used to identify the optimal number of clusters (k). The value of k=2. Figure 5 shows the result of the silhouette method.

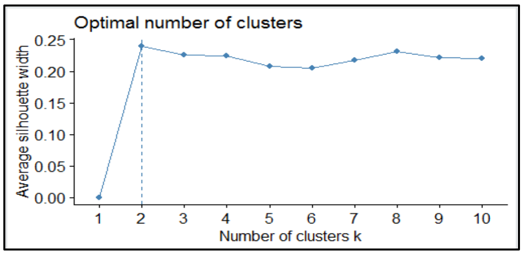


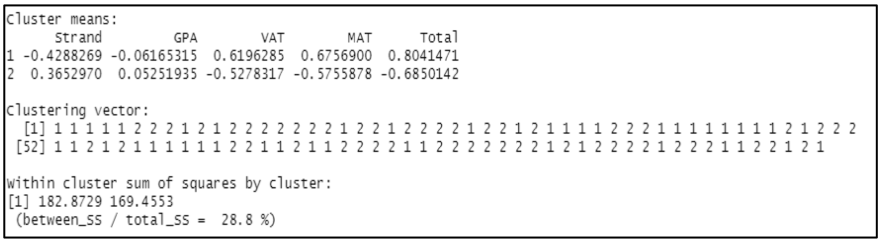
Figure 5. Optimal number of clusters taken silhouette method

These steps ensured the dataset was prepared and streamlined for cluster analysis, allowing for the identification of meaningful groups or clusters among the students.

4. RESULTS AND DISCUSSION

4.1. Cluster analysis using k-means

The dataset was clustered using k=2. Figure 6 shows the result of k-means clustering. The cluster plot for k-means applied on the dataset is shown on Figure 7. The admission dataset for academic year 2019-2020 was used to perform clustering of first year college students using the k-means algorithm, resulting in two clusters. These clusters based on the cluster plot can be: i) cluster 1-social science courses, putting together the education and psychology courses and ii) cluster 2-management courses, grouping together business management, hotel management, tourism management, and information technology. Alternatively, the clusters can also be viewed as: cluster 1 are board courses and cluster 2 are non-board courses. The findings suggest that students enrolled in social science classes to gain knowledge and become more aware of the world. Social science and education classes have an effect on the environment in which the students live. As a result, courses in education and social science were developed using k-means and placed in the same cluster. The expectations placed on the enrolled students to be able to gain knowledge and skills that will contribute to and influence the socialization of society are indicative of the commonalities between these courses' objectives and impact on society.



```
Cluster means:
      Strand         GPA        VAT        MAT      Total
1 -0.4288269 -0.06165315  0.6196285  0.6756900  0.8041471
2  0.3652970  0.05251935 -0.5278317 -0.5755878 -0.6850142

Clustering vector:
  [1] 1 1 1 1 1 2 2 2 1 2 1 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 1 2 2 1 2 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 2 1 2 2 2
 [52] 1 1 2 1 2 1 1 1 1 1 1 2 2 1 1 2 1 1 2 2 2 2 1 1 2 2 2 2 2 2 2 1 2 1 2 2 2 2 1 2 2 2 1 1 2 2 1 2 1

Within cluster sum of squares by cluster:
[1] 182.8729 169.4553
 (between_SS / total_SS =  28.8 %)
```

Figure 6. Result of k-means clustering

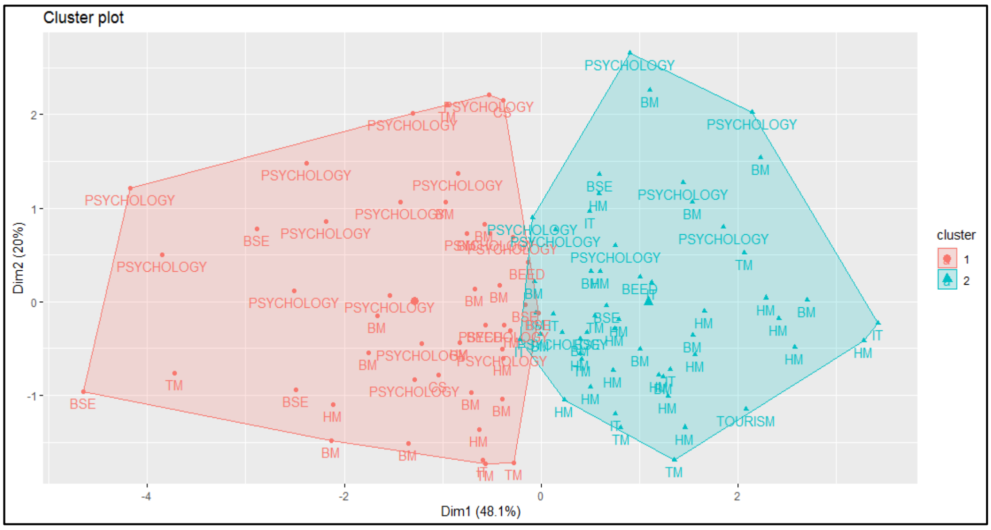


Figure 7. Cluster plot of k-means

Based on the similarity of their traits in the admission dataset, the clustering algorithm divided the first-year college students into two clusters. The qualities or traits that were employed for clustering form the basis of the final groupings. It appears that the clustering algorithm in this instance used the students' course preferences as the main feature for clustering. The algorithm divided the students who selected psychology and education into one cluster and the students who selected business management, hotel management, tourism management, and information technology into a different cluster based on the cluster plot. The resulting clusters can be interpreted as grouping students who have similar academic interests or career goals. For example, cluster 1 (social science courses) could be interpreted as students who are interested in pursuing careers in education or psychology, while cluster 2 (management courses) could be interpreted as students who are interested in business or technology-related careers. The clusters can also be thought of as putting students into groups based on whether or not they have selected board courses. Board courses often relate to courses needed to pass a certain board exam, like one in engineering or medicine. The two groups that were discovered may have been the result of the clustering method using the board courses as a characteristic. The admission dataset's combination of several variables that indicate students' interests, objectives, and academic backgrounds may be the cause of the clustering outcome overall. Depending on the dataset and the particular clustering technique employed, the specific features the clustering algorithm uses and the interpretation of the generated clusters may change.

4.2. Using hierarchical clustering

The dataset was clustered using k=2. Figure 8 shows the result of hierarchical clustering presented in a cluster dendrogram. The cluster plot for hierarchical clustering applied on the dataset is shown on Figure 9. Using hierarchichal clustering algorithm, it obtained with the same result as clustering using the k-means algorithm. The cluster 1 is composed of board courses, while the cluster 2 is composed of non-board courses. Academic courses that aim to improve students' similar prerequisites in pursuing and deepening their understanding of business are grouped together as business courses in cluster 2. Courses in this cluster share qualities with those expected of registrants, including standards that help them grow in terms of project management, leadership, critical, and strategic thinking.

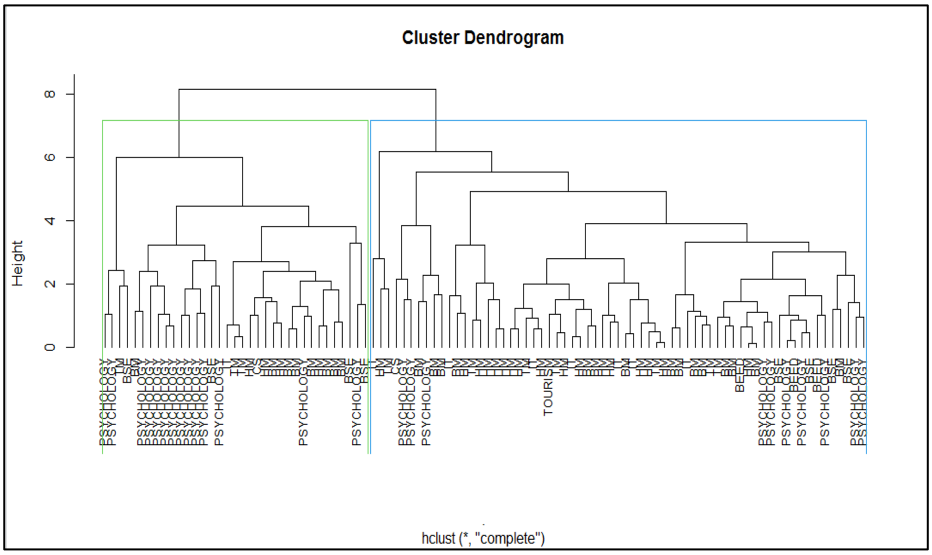


Figure 8. Cluster dendrogram

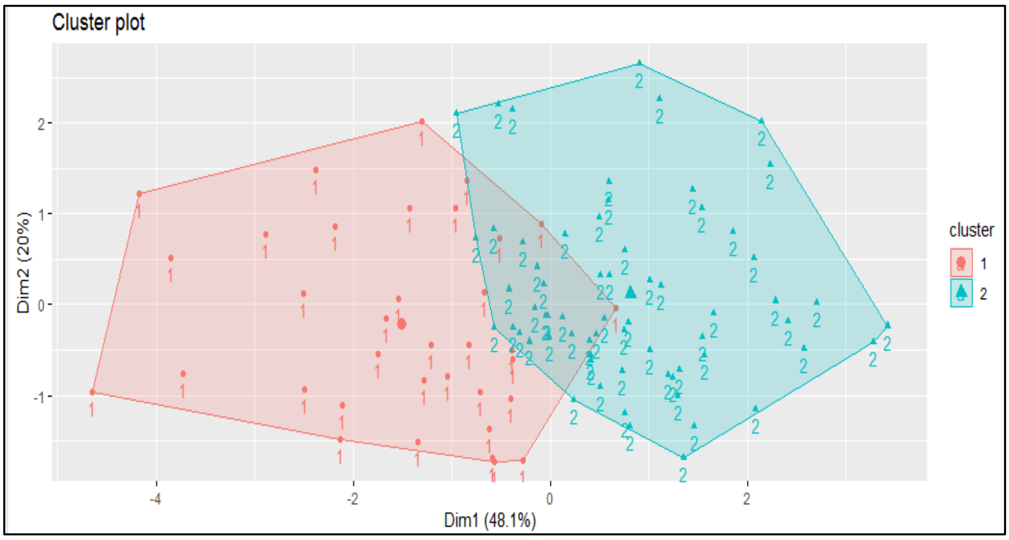
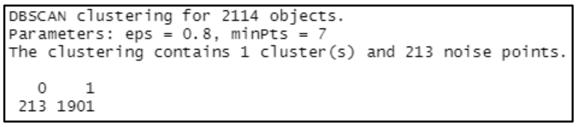


Figure 9. Cluster plot of hierarchical clustering

Like the k-means approach, the hierarchical clustering algorithm divided the first-year college students into two clusters based on the admission dataset's commonalities. By using the hierarchical clustering process, groups are iteratively merged or split according to how similar they are to one another. Based on the cluster descriptions produced, it appears that both the k-means method and the hierarchical clustering algorithm both used the students' course selection as a characteristic. The method put all board courses in cluster 1 and all non-board courses in cluster 2, respectively. Additionally, it appears from the cluster descriptions that the algorithm determined shared traits or features among the courses in cluster 2. Specifically, it implies that cluster 2 courses are designed to strengthen students' foundations for pursuing and extending their grasp of business. Moreover, the courses in this cluster share characteristics with those that are of expected enrollees, such as project management, leadership, and critical and strategic thinking. These interpretations of the clustering results could be based on a greater comprehension of the academic programs and courses offered by the college, as well as the professional objectives and interests of the students. It is possible that the clustering algorithm picked up on these tendencies in the admission dataset and grouped the students accordingly. It is important to note that these interpretations are based on the information provided and may not accurately reflect the underlying trends in the data.

4.3. Using density-based spatial clustering of applications with noise

The dataset was clustered using minPts=7. Figure 10 shows the result of k-means clustering. The cluster plot for k-means applied on the dataset is shown on Figure 11. As previously stated, both k-means and hierarchical clustering algorithms produced two clusters with similar cluster plots. In contrast, DBSCAN algorithm resulted in only one cluster. It indicates that there were not enough dense regions in the data for the system to identify multiple clusters. This could be due to several factors, such as the choice of density parameter, distance measure, or data quality. The clusters may not provide sufficient information to enable administrators to determine how to allocate courses. This is because these clusters are based solely on the courses that students choose to enroll in and do not take into account other factors such as students' academic backgrounds, interests, or career objectives.



```
DBSCAN clustering for 2114 objects.
Parameters: eps = 0.8, minPts = 7
The clustering contains 1 cluster(s) and 213 noise points.

   0    1
 213 1901
```

Figure 10. Result of DBSCAN

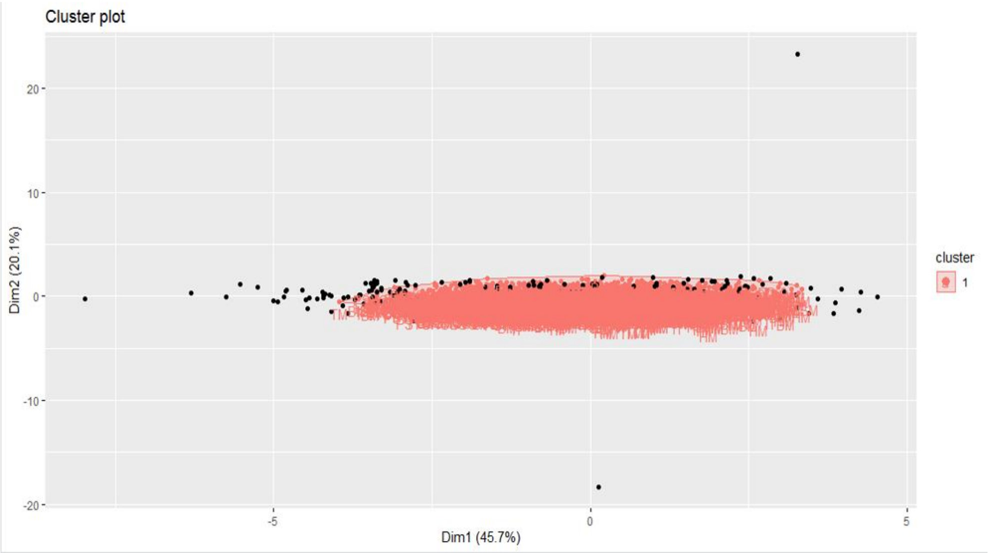


Figure 11. Cluster plot of DBSCAN

Using clustering as a decision-support tool to guide students' course work is still a good idea. Clustering can be a useful method for identifying patterns and similarities in large datasets, which can then be used to make informed decisions about course offerings and student assignments. To ensure that the clustering results are meaningful and practical, administrators can consider incorporating other factors and data sources into the clustering algorithm, such as statistics on student performance or survey findings.

The clustered results for board and non-board courses offer a vague indication for choosing the subjects that will be assigned to students. If this takes place, the university's courses will lose their significance and distinctiveness. To give the school's name some value, administrators should consider using this as a decision-support tool in directing students' course work.

5. CONCLUSION

Using the different clustering algorithms, in particular, k-means, hierarchical, and DBSCAN, it yielded an almost similar number of clusters and cluster plots. Its implementation in R showed that the admission dataset for first year college students group can be divided into 2 clusters. In this light, cluster analysis may be able to give insights to admission officers as to where students will be placed. The algorithms utilized in this work are effective resources for studying data on student admissions. The schools can make better decisions regarding the admissions process and determine the traits of successful student-applicants by spotting patterns and trends in the data. The result of the study can aid other HEI in making more efficient judgments in enhancing their existing procedures and increasing the bar for students' admission, assistance, and support. It is helpful in predicting students' turn over to board examinations, retention, increasing graduation rates, curriculum revisions and enhancement, and effectively assess the performance of the university.

Using the clustering methodology, admissions officers can classify prospective students into clusters according to their academic qualifications and personal interests. Informed judgments about course placement and counseling can be made by admissions staff using this data, which will eventually increase student success and retention. To enhance and change the curriculum, patterns and trends in the courses that students in each cluster choose to take can be discovered using the clustering analysis. With the information provided in the result of this study, the curriculum of educational institutions may be improved to better address the needs and interests of the students. The institution may evaluate the efficacy of its programs and policies and make data-driven decisions for improvement by following changes in the clustering patterns over time. Overall, it offers insightful information on the requirements and characteristics of university students. The institution may boost student achievement, improve operations, and improve its reputation and value by leveraging these insights to guide decision-making. For future works, other attributes such as locality, family income, and skills may also be looked into.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Cavite State University for supporting this undertaking. This work was funded with a grant number PHDR-029A-2023.

REFERENCES

[1] M. I. P. Conchada and M. M. Tiongco, “A Review of the Accreditation System for Philippine Higher Education Institutions,” Makati, 2015.

[2] J. A. Esquivel and J. A. Esquivel, “A Machine Learning Based DSS in Predicting Undergraduate Freshmen Enrolment in a Philippine University,” *International Journal of Computer Trends and Technology*, vol. 69, no. 5, pp. 50–54, 2021, doi: 10.14445/22312803/ijctt-v69i5p107.

[3] J. Luan, “Data Mining Applications in Higher Education,” *SPSS Executive*, vol. 7, pp. 1–8, 2004, doi: 10.1201/b15783-13.

[4] K. N. Shah, M. R. Patel, N. V Trivedi, P. N. Gadariya, R. H. Shah, and N. Adhvaryu, “Study of Data Mining in Higher Education-A Review,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 1, pp. 455–458, 2015.

[5] N. Khan, “Decision making Assessment model for University admission,” *International Journal of Advances in Computer Science and Technology*, vol. 11, no. 7, pp. 23–28, 2022, doi: 10.30534/ijacst/2022/011172022.

[6] E. Daniati, “Decision Support Systems to Determining Programme for Students Using DBSCAN And Naive Bayes: Case Study: Engineering Faculty Of Universitas Nusantara PGRI Kediri,” in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019, pp. 238–243, doi: 10.1109/ICAIIT.2019.8834474.

[7] M. I. A.-Twijri and A. Y. Noaman, “A New Data Mining Model Adopted for Higher Institutions,” *Procedia Computer Science*, vol. 65, pp. 836–844, 2015, doi: 10.1016/j.procs.2015.09.037.

[8] H. A. Mengash, “Using data mining techniques to predict student performance to support decision making in university admission systems,” *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

[9] X.-F. Lei, M. Yang, and Y. Cai, “Educational data mining for decision-making: a framework based on student development theory,” in *Proceedings of the 2nd Annual International Conference on Electronics, Electrical Engineering and Information Science (EEEIS 2016)*, 2017, pp. 628–641, doi: 10.2991/eeeis-16.2017.76.

[10] A. F. Mashat, M. M. Fouad, P. S. Yu, and T. F. Gharib, “Efficient Clustering Technique for University Admission Data,” *International Journal of Computer Applications*, vol. 45, no. 23, pp. 39–42, 2012.

[11] H. Muttaqien, M. Lutfi, M. KH, A. Muis, and H. Zainuddin, “Recommendation of Student Admission Priorities Using K-Means Clustering,” in *ICOST 2019: 1st International Conference on Science and Technology*, Gent, Belgium: European Alliance for Innovation, 2019, pp. 375–382, doi: 10.4108/eai.2-5-2019.2284614.

[12] T. M. Facca and S. J. Allen, “Using Cluster Analysis to Segment Students Based on Self-Reported Emotionally Intelligent Leadership Behaviors,” *Journal of Leadership Education*, vol. 10, no. 2, pp. 72–96, 2011.

[13] M. Hickendorff, W. J. Heiser, C. M. v. Putten, and N. D. Verhelst, “Clustering Nominal data with Equivalent Categories,” *Behaviormetrika*, vol. 35, no. 1, pp. 35–54, 2008, doi: 10.2333/bhmk.35.35.

[14] N. K. Nissa, “Clustering Method using K-Means, Hierarchical and DBSCAN (using Python),” *Medium*, 2020. [Online]. Available: https://nzlul.medium.com/clustering-method-using-k-means-hierarchical-and-dbscan-using-python-5ca5721bbfc3 (accessed Mar. 22, 2022).
[15] O. R. Battaglia, B. D. Paola, and C. Fazio, “A New Approach to Investigate Students’ Behavior by Using Cluster Analysis as an Unsupervised Methodology in the Field of Education,” *Applied Mathematics*, vol. 07, no. 15, pp. 1649–1673, 2016, doi: 10.4236/am.2016.715142.
[16] D. Deng, “DBSCAN Clustering Algorithm Based on Density,” in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949–953, doi: 10.1109/IFEEA51475.2020.00199.
[17] J. O. Oyelade, O. O. Oladipupo, and I. C. Obagbuwa, “Application of k-Means Clustering algorithm for prediction of Students’ Academic Performance,” *International Journal of Computer Science and Information Security*, vol. 7, no. 1, pp. 292-295, 2010.
[18] G. Santosa, Y. Lukito, and A. Rachmat, “Classification and Prediction of Students’ GPA Using KMeans Clustering Algorithm to Assist Student Admission Process,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 7, no. 1, pp. 1-10, 2021, doi:10.20473/jisebi.7.1.1-10.
[19] D. S. Maylawati, T. Priatna, H. Sugilar and M. A. Ramdhani, “Data science for digital culture improvement in higher education using K-means clustering and text analytics, ” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, pp. 4569-4580, 2020, doi: 10.11591/ijece.v10i5.pp4569-4580.
[20] M. Iqbal, M. B. Ryando, T. Triono and N. Nurmaesah, “Clustering of Prospective New Students using Agglomerative Hierarchical Cluserting,” *Proceeding International Conference on Information Technology, Multimedia, Architecture, Design, and E-Business*, 2022, vol. 2, pp. 183-192.
[21] M. Goyal and R. Vohra, “Applications of Data Mining in Higher Education,” *International Journal of Computer Science Issues*, vol. 9, no. 2. pp. 113-120, 2012.
[22] A. J. Bowers, “Analyzing the ongitudinal K Analyzing the longitudinal K-12 grading hist ading histories of entir ories of entire cohor e cohorts of students: Grades, data driven decision making, dropping out and hierarchical cluster analysis,” *Practical Assessment, Research and Evaluation*, vol. 15, pp. 1-18, 2010, doi: 10.7275/r4zq-9c31.
[23] E. Daniati, “Decision Support Systems to Determining Programme for Students Using DBSCAN And Naive Bayes: Case Study: Engineering Faculty Of Universitas Nusantara PGRI Kediri,” *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, Yogyakarta, Indonesia, 2019, pp. 238-243, doi: 10.1109/ICAIIT.2019.8834474.
[24] A. F. M. Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman and M. Aliff, "Clustering Analysis for Classifying Student Academic Performance in Higher Education,” *Applied Science*, vol. 12, no. 19, pp. 1-22, 2022, doi: 10.3390/app12199467.
[25] N. Valarmathy, S. Krishnaveni, “Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining,” *International Journal of Recent Technology and Engineering*, vol. 7, no. 655, pp. 103-113, 2019.
[26] Z. Gniazdowski and M. Grabowski, “Numerical Coding of Nominal Data,” *Zeszyty Naukowe WWSI*, vol. 12, no. 9, pp. 53–61, 2015, doi: 10.26348/znwwsi.12.53.

BIOGRAPHIES OF AUTHORS

Erwin Lanceta Cahapin received the industrial technology degree from Rogationist College, Silang Cavite, Philippines in 2010. He received the master of science degree in Engineering Education Major in Computer Engineering from Rizal Technological University Mandaluyong, Philippines in 2020. He is pursuing his doctor in information technology degree at Technological Institute of the Philippines in Manila, Philippines. Currently, he is a faculty member under the Department of Information Technology, and the Program Coordinator of Computer Science, Cavite State University–Silang Campus, Silang, Cavite, Philippines. His research interests include data mining, natural language processing, and artificial intelligence. He can be contacted at email: elcahapin@cvsu.edu.ph.

Beverly Ambagan Malabag graduated from Cavite State University in Indang, Cavite Philippines, with a bachelor of science in Computer Engineering in 2007 and a master of engineering with a major in Computer Engineering in 2012. She is pursuing her doctor of engineering degree at the Technological Institute of the Philippines in Quezon City, Philippines with specialization in computer engineering. She has been a faculty member since 2007 and is presently the chairperson of the Department of Information Technology at Cavite State University–Silang Campus, Silang, Cavite, Philippines. Image processing, artificial intelligence, and data mining are some of her research interests. She can be contacted at email: beverlymalabag@cvsu.edu.ph.

Cereneo Sailog Santiago Jr. received his bachelor of science in Computer Science, master of arts in Education, and master in Information Technology in 2007, 2018 and 2021, respectively. Currently serving as the coordinator of the Knowledge Management Unit, Program Coordinator for the BS in Information Technology, head of the office of Student Affairs and Services, and faculty member of the Department of Information Technology at Cavite State University-Silang Campus, Silang, Cavite, Philippines. Prior to his current roles, he also served as the coordinator for research and extension of the campus. His research interests are related to ICT education, online learning, computing and technologies, developmental, action, and institutional research. He is pursuing his doctorate degree in Research and Evaluation. He can be contacted at email: cssantiago@cvsu.edu.ph.

Jocelyn L. Reyes is an assistant professor V of Cavite State University where she has been teaching physics, mathematics, physical science, and other related fields since 1982. She obtained her bachelor of science in physics for teachers at Philippine Normal College in consortium with De La Salle University under PNC-NSBD project 7405 Ed. She pursued completed all the academic requirements for Ph.D. Physics, again at DLSU Manila. She graduated at St. Jude College Manila where she took Master of Arts in Education (MAEd), major in science. She obtained her degree in Ph.D. Educational Management, also at SJC. Affiliated with the Department of Physical Science under the College of Arts and Sciences, she is an active member of the teaching force which supports the various curricular offerings of the different colleges at CvSU main campus. Spending almost 32 years in her chosen career, she has served various positions in the academe, namely: head, socio-cultural affairs (SOSCA) under the office of student affairs; head, higher ed-instructional materials and development unit (IMDU); college research development and extension (RD&E) coordinator, CAS; Department RD&E coordinator and Department Chair, both DPS, among others. At present, she is the campus administrator of Cavite State University-Silang Campus. She is an author, co-author, editor, and coordinator of different instructional materials in math and science. She can be contacted at email: jocelynreyes@cvsu.edu.ph.

Gemma S. Legaspi is an assistant professor II of Cavite State University–Don Severino De Las Alas Campus, Indang, Cavite. She graduated cumlaude from the same institution, formerly Don Severino Agricultural College, with a degree in bachelor of secondary education major in mathematics. She finished her master of arts in education major in mathematics at St. Jude College, Manila. At present, she is taking doctor of philosphy in Educational Management, also at St. Jude College. She took up units in doctor of philosophy in mathematics–straight program at De La Salle University, Manila. She is a member of mathematical society of the Philippines (MSP), Philippine council of mathematics educators, Inc (MATHTED), mathematic teachers association of the Philippines (MTAP), mentors and leaders society of the Philippines (MLSP), and State Universities and Colleges Teachers Educators Association (SUCTEA). She is an author, co-author, editor, and coordinator of the different instructional materials in math and science. She can be contacted at email: gemmalegaspi@cvsu.edu.ph.

Karl Louise Adrales received his bachelor of science in International Studies major in European Studies in 2019 and currently taking master of arts in foreign service at Lyceum of the Philippines University Manila. He is currently working as administrative assistant and a part time instructor at Cavite State University-Silang Campus. He can be contacted at email: karllouise.adrales@cvsu.edu.ph.